Audio Engineering Society

Convention Paper 8086

Presented at the 128th Convention
2010 May 22–25 London, UK

Integrating musicological knowledge into a probabilistic framework for chord and key extraction

Johan Pauwels[1], and Jean-Pierre Martens[1]

[1] *Department of Electronics and Information Systems (ELIS), Ghent University, Belgium*

Correspondence should be addressed to Johan Pauwels (johan.pauwels@elis.ugent.be)

ABSTRACT

In this contribution a formerly developed probabilistic framework for the simultaneous detection of chords and keys in polyphonic audio is further extended and validated. The system behaviour is controlled by a small set of carefully defined free parameters. This has permitted us to conduct an experimental study which sheds a new light on the importance of musicological knowledge in the context of chord extraction. Some of the obtained results are at least surprising and, to our knowledge, never reported as such before.

1. INTRODUCTION

In Western tonal music, a chord is defined as a fixed collection of simultaneously sounding notes with an associated name. Because chords provide the harmonic backbone over which the melody is played, they are the basic building blocks of a song. Multiple chords in sequence establish a broader musical context, called a key. Since chords can be interpreted in a key, the latter imposes certain expectations on its constituting chords.

Chord extraction is the process of converting an audio recording into a stream of chord symbols. The resulting chord sequence can be used directly, for instance to learn how to play a particular song. It can also be used as an intermediate representation for a variety of indirect applications, for instance to generate an automatic accompaniment or to query for similar songs in a database.

Traditionally, chord extraction is done by trained musical experts. This makes collecting a large amount of chord transcriptions laborious, while certain applications like similarity querying are only useful when a large database is available. Automated chord transcription would be very welcome

in such cases and thus poses an actively researched challenge.

Any chord extraction system consists of an acoustic front-end converting the audio signal into a stream of acoustic feature vectors and a back-end converting this acoustic feature stream into a timed chord-label sequence. The acoustic feature vectors are usually chroma profiles emerging from a spectral analysis of subsequent frames. Based on whether or not the back-end incorporates a form of prior musicological knowledge, previous work can be split into 2 categories. Back-ends without musical knowledge usually perform a frame-by-frame chord classification followed by some post-processing [1, 2, 3]. Back-ends including musicological knowledge conduct an integrated search for the most likely state sequence through a finite state automaton, usually a Hidden Markov Model (HMM) [4, 5, 6, 7]. Our approach belongs to the last category and it permits us to include musicological knowledge with different weightings. This way we could thoroughly investigate the impact of the musical knowledge on the extraction accuracy.

While we are not the first to attempt the simultaneous extraction of keys and chords [6, 8], we believe to be the first to do this with a musicological model in terms of relative chords in a key, only depending on the key mode. This approach permits us to exploit the parallelism between keys differing in mode but not in tonic and it also supports distinct idiomatic chord sequences per mode.

With 4 recognizable chord types, our system (like that of [2]) tries to pursue a good balance between the ambitious number of chord types distinguished in early work [1, 7] and the conservative major–minor distinction made in later work [3, 4, 5, 6], work that was influenced by the evaluation protocol adopted in the MIREX contest.

2. OVERVIEW OF THE SYSTEM

The input audio file is supplied to a three-stage front-end. In the first stage, the waveform is resampled to 8 kHz and converted to mono. The resulting waveform is then subjected to a spectral analysis with the following characteristics: the analysis is performed on 150 ms long fragments (called frames), the frames are weighted with a Hamming window and for each windowed frame, a chroma profile is calculated. The frame shift, defined as the time between subsequent frames, is 20 ms. The local spectral analysis is followed by an integration stage which computes, at multiples of the *frame shift*, the mean of the chroma profiles observed in a specified number of subsequent frames. The outputs of the front-end are represented by the acoustic observation vector sequence $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_N\}$. The time shift between subsequent vectors is hereafter called the *hop size* whereas the *segment size* refers to the length of the integration interval. Both are expressed as multiples of the frame shift.

The back-end performs an integrated dynamic programming search for the most likely alignment of the acoustic observation sequence with the states of a finite state machine. The latter consists of 24 x 48 states, each representing a key-chord pair. For the keys, we discern two modes — major and minor — and 12 possible tonics for each mode. For the chords, we discern four types of triads — major, minor, diminished and augmented — and 12 possible roots. The simultaneous determination of key and chord labels offers us a nice opportunity to deal with key modulations in a natural way. Nevertheless, the results discussed in this paper uniquely refer to the chord extraction properties of our system.

2.1. Chroma profile analysis

As in many other systems, our acoustic observation vectors represent chroma profiles. However, the calculation of these profiles differs from what is commonly used. In its simplest form, a chroma profile is just a log-frequency representation of the spectral content, folded into a single octave. The problem with such a representation is that e.g. the third harmonic of a pitch folds into a chroma that is located at $+7$ or -5 semitones with respect to the fundamental. Consequently, it will add evidence to a second pitch class that is not necessarily present as a played note in the signal. We therefore developed a more complex implementation [9] which aims at maximally coupling the higher harmonics to their fundamental frequency. To that end, it uses multiple pitch tracking techniques. Ideally, if the harmonic-to-fundamental coupling was perfect, the chroma profile would only represent notes that are actually played, and the chord extraction would boil down to a simple pattern matching operation. Note

that our approach of dealing with higher harmonics in the front-end is orthogonal to the approach advocated in [3] and [5], where the harmonics are accounted for in the back-end.

Since fundamental frequencies below 100 Hz usually represent bass-lines and therefore rarely contribute unique chromas to a chord, only fundamental frequencies larger than 100 Hz are considered during the chroma analysis. Because a fundamental frequency needs to be supported by at least one harmonic, it follows that the highest detectable fundamental frequency is 2000 Hz.

To reduce the sensitivity to the sound volume, the values of the non-zero chroma profiles are rescaled by dividing them by the sum of all components.

An implementation of our chroma profile analysis is publicly available in MARSYAS [10], but on request it can also be supplied as a stand-alone executable or a Matlab program.

2.2. Integrating chroma profiles

The integration of chroma profile patterns over time can in principle be performed in two ways: either by taking the mean over fixed length intervals, or by taking the mean over variable length intervals which are presumed to reveal an interesting segmental context (e.g. intervals between subsequent beats). In the latter case, one observation vector per integration interval is generated, meaning that these vectors represent subsequent (disjoint) intervals. In the former case, the hop size is fixed and can range from 1 to S, the segment size. In this paper, we only consider this case because it is not burdened by errors made by some higher-level segmentation algorithm. By varying the hop size, we can trade off computational load for time resolution, and possibly, accuracy.

2.3. Probabilistic framework

The back-end implements a unified probabilistic framework for the simultaneous recognition of chords and keys. It builds upon former work that was published in [11].

The objective of the back-end is to retrieve the most likely sequence of states $\hat{Q} = \{\hat{q}_1, \ldots, \hat{q}_N\}$ for the acoustic observation sequence $\mathbf{X}$ of length N. Each state q consists of the combination of a key k and a chord c, and the state sequence is equivalent to a sequence of key and chord labels: in what follows, q_n is equivalent to (k_n, c_n). Using Bayes's rule, it follows that

$$\begin{aligned}\hat{Q} &= \arg\max_Q P(Q|\mathbf{X}) \\ &= \arg\max_Q P(Q, \mathbf{X}) \\ &= \arg\max_Q P(Q)\, P(\mathbf{X}|Q)\end{aligned}$$

The acoustic likelihood $P(\mathbf{X}|Q)$ is calculated by means of an acoustic model. By making the standard assumption that acoustic observations emitted in the same state are independent of each other, the acoustic likelihood can be factorised as follows

$$P(\mathbf{X}|Q) = \prod_{n=1}^{N} P(\mathbf{x}_n|k_n, c_n)$$

Since the same chroma profile can occur in all keys, one can assume that the acoustic observation vector $\mathbf{x}_n$ only depends on the chord label c_n. Consequently the acoustic likelihood can be computed as

$$P(\mathbf{X}|Q) = \prod_{n=1}^{N} P(x_n|c_n)$$

The prior probability $P(Q)$ is computed by making the first order Markov assumption. This means that

$$P(Q) = \prod_{n=1}^{N} P(q_n|q_{n-1})$$

To model that a state will be visited for a variable number of frames, a self-transition is attached to each state. The set of self-transition probabilities constitute a chord duration model. We opted for a very simple geometric model, represented by a single $P_s = P(q_n = q_{n-1})$ for all states.

The transitions between states are governed by some musicological stochastic bigram model in terms of keys and chords. It can be decomposed into a key transition model and a chord transition model:

$$P(K, C) = \prod_{\substack{n=1 \\ q_n \neq q_{n-1}}}^{N} P(k_n|k_{n-1}, c_{n-1})\, P(c_n|k_n, k_{n-1}, c_{n-1})$$

We now argue that in first approximation, the previous chord c_{n-1} has only a negligible impact on the identity of the current key label. Furthermore, in accordance with common practice in harmonic analysis we always interpret chord c_n in the context of the already hypothesized key k_{n-1}. In fact, if c_n is a pivot chord that can be interpreted as the end chord of a harmonic sequence in key k_{n-1}, but also as the start of a harmonic sequence in a different key k_n, it is musically intuitive to interpret the transition between c_{n-1} an c_n as a chord transition in key k_{n-1} and the transition between c_n and c_{n+1} as a transition in key k_n. Based on these arguments, the musicological model can be simplified to

$$P(K,C) = \prod_{\substack{n=1 \\ q_n \neq q_{n-1}}}^{N} P(k_n|k_{n-1})\, P(c_n|k_{n-1}, c_{n-1})$$

To have more control over the importance of the musicological model in the system and over the relative importance of its key and the chord transition components, we introduced two extra control parameters τ and κ. We also work in the log_{10}-domain, so that

$$\hat{Q} = \arg\max_{Q} \sum_{n=1}^{N} \big[\log P(x_n|c_n) + \log P(q_n|q_{n-1})\big]$$

$$\begin{aligned} \log P(q_n|q_{n-1}) &= \log(P_s) && (q_n = q_{n-1}) \\ &= \log(1-P_s) + \tau\mathfrak{L}_M && (q_n \neq q_{n-1}) \end{aligned}$$

$$\begin{aligned} \mathfrak{L}_M = \big[&\kappa \log P(k_n|k_{n-1}) \\ &+ (1-\kappa)\log P(c_n|k_{n-1}, c_{n-1})\big] \end{aligned}$$

2.4. Two acoustic model approaches

The acoustic model calculates the likelihood of an acoustic observation vector given a proposed chord. We consider two simple approaches which both rely on the definition of a 12-dimensional binary template $\mathbf{t}(c)$ to represent the chord c. The m-th element $t_m(c)$ of $\mathbf{t}(c)$ is 1 if chord c implies a component of chroma m, and 0 in the other case. Since we stick to triads, all considered chords will have a template with three non-zero values.

The first approach is to consider the cosine similarity between the observation vector and the chord template as the acoustic likelihood. Since the elements of the observation vectors and the chord templates are positive numbers, the cosine similarity ranges from 0 (no evidence found for any of the chromas present in the chord) to 1 (the observation vector has and only has contributions on the chromas present in the chord).

The second approach is the one adopted in the original version of our system [11]. It assumes that the elements of the observation vector can be considered independently of each other, and that the acoustic likelihood can thus be factorised:

$$P(x_n|c_n) = \prod_{m=1}^{12} P(x_{nm}|t_m(c_n))$$

Since $t_m(c)$ is either 1 or 0, there are only two probability distributions to model. A simple model is the following

$$\begin{aligned} P(x|0) &= \nu_0 \left(P_0 + (1-P_0)e^{-\frac{x^2}{2\sigma^2}}\right) && x \in (0,1) \\ P(x|1) &= \nu_1 \left(P_0 + (1-P_0)e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) && x \in (0,\mu) \\ &= \nu_1 && x \in (\mu,1) \end{aligned}$$

The quantities ν_0 and ν_1 are normalisation factors whereas P_0 is an offset which preserves some probability density at $x = 0$ or 1 for a template value of 1 and 0 respectively. Since we assume triads, and since the observation vectors are normalised, it follows that $\mu = 0.33$ is an appropriate choice. In order to obtain sufficiently discriminating distributions we chose $\sigma = 0.13$ (sufficiently smaller than μ). The parameter P_0 was kept as a control parameter.

2.5. A simple key transition model

The probability that after the termination of a chord, the system will move to a chord in the same key is controlled by a parameter P_k. The probability of moving to another key is derived from the Lerdahl distance [12] between the two keys involved. If $d(k_n, k_{n-1})$ is the Lerdahl distance between the diatonic chord on the tonic of the destination key (k_n) and the diatonic chord on the tonic of source key (k_{n-1}), then

$$\begin{aligned} P(k_n|k_{n-1}) &= P_k && k_n = k_{n-1} \\ &= \nu_k e^{-\frac{d(k_n,k_{n-1})}{d_k}} && k_n \neq k_{n-1} \end{aligned}$$

The factor ν_k is a normalisation factor to ensure that

$$\forall Y : \sum_{\substack{X \\ X \neq Y}} P(k_n = X | k_{n-1} = Y) = 1 - P_k$$

and d_k is the mean distance between two keys.

Since there are only a few key modulations in our dataset, it would have been very difficult to demonstrate the impact of the key transition model. Therefore, P_k was fixed to 0.99 throughout our experiments, and no alternatives to the Lerdahl-based model were investigated.

2.6. A relative chord transition model

Because the chord duration has already been modelled by the term P_s, the musicological model only has to model the transitions between different states. In contrast to all other systems we know of, we express the chord transition probabilities $P(c_n|k_{n-1}, c_{n-1})$ in terms of chord degrees and key modes. In particular, we express chords c_n and c_{n-1} as degrees in key k_{n-1}, and we denote these relative chords as c'_n and c'_{n-1}. Then we construct a model in which the envisaged chord transition probability only depends on the mode m_{n-1} of key k_{n-1} and on the relative chord transitions. This way we exploit the parallelism between keys that differ in tonic, but not in mode (as in [6], but avoiding the explicit construction of parallel systems for every key). Our approach permits us to construct distinct transition models for major and minor keys, for which distinct idiomatic chord sequences exist. Other systems such as [4, 5, 7] do not simultaneously extract the key, thus preventing this distinction to be made. Thanks to the above simplification there are considerably fewer transitions to model, and consequently, one can harvest many more training examples for each of them in a training corpus of a restricted size.

We assessed three different models for the outgoing transitions of a state: a uniform distribution assigning equal probabilities to all outgoing transitions, a bigram model trained on annotated data and a theoretical model based on the Lerdahl distance.

In the theoretical model, the probability mass is divided into a part P_d that is available for transitions to a degree that is diatonic in m_{n-1} and a part $1 - P_d$ that is available for the other transitions. The transitions between two degrees which are both diatonic in m_{n-1} are on their turn distributed according to the Lerdahl distance between chords in the same key. The non-diatonic transitions are uniformly distributed. If $d(c'_n, c'_{n-1})$ represents the Lerdahl distance between two chords in the same key, the relative chord transition model can be formulated as

$$\begin{aligned} &P(c'_n|m_{n-1}, c'_{n-1}) \\ &\quad = \nu_c \, P_d \, e^{-\frac{d(c'_n, c'_{n-1})}{d_c}} \\ &\qquad (c'_n, c'_{n-1} \in \mathfrak{D}(m_{n-1})) \\ &\quad = \frac{P_d}{N_d} \\ &\qquad (c'_n \in \mathfrak{D}(m_{n-1}) \text{ and } c'_{n-1} \notin \mathfrak{D}(m_{n-1})) \\ &\quad = \frac{1 - P_d}{N_{nd}} \\ &\qquad (c'_n \notin \mathfrak{D}(m_{n-1})) \end{aligned}$$

where $\mathfrak{D}(m)$ represents the set of chord degrees diatonic in mode m. As before, ν_c is a norm factor such that for all combinations of Y and Z

$$\sum_X P(c'_n = X | m_{n-1} = Y, c'_{n-1} = Z) = 1$$

The quantities N_d and N_{nd} represent the number of diatonic, respectively non-diatonic chords for mode Y. The factor d_c is the mean Lerdahl distance between two diatonic chords in a key. By setting P_d to 1, one can ensure that only diatonic chords will be generated.

For creating the trained model, we counted chord transitions in a development dataset composed of 142 manually annotated 30 s excerpts from musical pieces covering a variety of tempi and genres. These counts were then converted to conditional bigram probabilities, with a small probability being held out for unseen transitions.

3. EVALUATION PROCEDURE

As an evaluation measure, we calculate the average overlap score (AOS) formerly introduced in the MIREX [13] contest. It is defined as the percentage of time the computed chord is equal to the annotated chord. To avoid the need for a mapping of complex chords to triads and a set of associated scoring rules, we only evaluate at times where one of the basic triads has been annotated. This leaves us with 2656 s (instead of the maximum 142 x 30 = 4260 s) of data.

	cosine		Gaussian	
S	$hop = 1$	$hop = S$	$hop = 1$	$hop = S$
1	41.58	41.58	41.61	41.61
3	46.86	46.80	46.82	46.76
5	49.84	49.71	49.81	49.72
11	55.47	54.97	55.38	54.89
21	60.86	59.98	60.74	59.86
31	63.27	60.86	63.19	60.86
41	64.75	62.58	64.62	62.39
51	65.75	62.03	65.71	61.94
71	66.33	60.48	66.29	60.39
91	65.57	60.02	65.52	60.07

Table 1: The AOS for different acoustic-only systems. Results are presented for different values of the hop size (in frames) and the segment size (in frames)

4. EXPERIMENTAL RESULTS

The experimental evaluation is intended to find optimal values for the free parameters P_s, P_d, τ and κ, and to demonstrate the impact of the different model components on the system accuracy.

4.1. Influence of the segment size on an acoustic-only system

In a first experiment, we evaluated a system using only acoustic models ($\tau = 0$). This system returns the chord sequence (without a key) which is most likely on the basis of acoustic information only. The results for the two acoustic models outlined in Section 2.4 are displayed in Table 1 for different values of the hop size and the segment size.

Averaging the observations clearly has a positive effect on the AOS. For a hop size of 1, the AOS increases until $S = 71$. This segment size corresponds to 1.42 seconds, which is very close to the average chord duration of 1.39 seconds found in the dataset. For longer segment sizes the drawback of masking boundary information outweighs the advantage of having a more stable input. For the case of disjoint segments (hop $= S$) the optimal point is significantly lower and located at $S = 41$. As soon as $S > 11$, the loss of time resolution comes into play as an additional detrimental factor.

Although the optimal result is attained for a relatively long segment size, we feel that for such a value an unacceptably large number of short chords will remain undetected. This problem will be solved by the addition of the chord duration model.

4.2. Importance of the distinct model components

In order to find the optimal balance τ between the acoustic and the musicological model, we fix κ to 0.5, and we search for sensible values of P_s and P_d which are easy to retrieve from a small annotated dataset.

Obviously, P_s depends on the hop size because it implies a mean chord duration of $P_s/(1 - P_s)$ hops. Since we found that the mean chord duration in our development set was about 1.4 seconds, we determined P_s as 1.4 divided by 1.4 plus the hop size in seconds. For determining P_d we just counted the fraction of chord changes to a diatonic chord with respect to all chord changes.

Since Table 1 showed little effects of changing the time resolution from 1 to S for segment sizes up to 11, we chose this frame size as our point of departure. Table 2 displays the optimal τ and associated AOS for all combinations of the two acoustic models and the three musicological models, and this for the two considered hop sizes.

Expectedly, the value of τ depends on the choice of the acoustic model and on the hop size. The latter is obvious since the number of between-state transitions (having a musicological transition) versus the number of self transitions (having an acoustic model probability) directly depends on the hop size.

A first important finding is that the attainable accuracy is very comparable for the two acoustic models, but with a small preference for the Gaussian model. The most striking results however are (1) that the uniform musicological model performs almost as well as the trained musicological model (the difference is less than 2%), and (2) that the trained model does not outperform the theoretical model.

Comparing these results to the ones in Table 1 lets us conclude that a duration model is indispensable for attaining good results, whereas a musicological model is not bringing a large additional accuracy gain.

	cosine		Gaussian	
	$hop = 1$	$hop = S$	$hop = 1$	$hop = S$
trained	$\tau = 2.3$ $AOS = 71.82$	$\tau = 0.32$ $AOS = 72.10$	$\tau = 14$ $AOS = 73.34$	$\tau = 1.9$ $AOS = 72.55$
theoretical $P_d = 0.80$	$\tau = 2.2$ $AOS = 72.14$	$\tau = 0.26$ $AOS = 71.72$	$\tau = 16$ $AOS = 74.10$	$\tau = 2.2$ $AOS = 72.65$
uniform	$\tau = 2.3$ $AOS = 70.56$	$\tau = 0.26$ $AOS = 70.05$	$\tau = 13$ $AOS = 71.95$	$\tau = 1.4$ $AOS = 70.96$

Table 2: Optimal τ with associated AOS for all combinations of acoustic and musicological models

4.3. Sensitivity to P_s and P_d

In view of the generalisation of results to new datasets with possibly other characteristics, we investigated how sensitive the results are to a change in the parameters P_s and P_d. We performed the experiments with the theoretical musicological model, a hop size of S and a segment size of 11. The results are displayed in Table 3a and Table 3b, respectively. Apparently, the sensitivity is quite low, especially for P_d.

4.4. Influence of the balance between key and relative chord transition model

Although only one key model was proposed, it is still possible to examine the effect of including it or not. By resetting P_s and P_d to their initial values and by changing κ from 0 to 0.75, we obtained the results listed in Table 3c. It appears that the initially assumed value of 0.5 is optimal, and that the overall gain in AOS attained by including a key transition model amounts to about 3%.

4.5. Influence of segment size for the complete system

Thus far, the complete system was only tested with a segment size of 11 frame shifts. We therefore tested our best systems (the two acoustic models each in combination with the theoretical musicological model, a hop size of 1 and optimal values for the other free parameters) for different values of S. The corresponding AOS values are shown in Table 4. We can conclude that the introduction of the duration model allows us to get rid of the long integration time.

4.6. Validation on the Beatles set

Finally, we performed a validation experiment on the wide-spread Beatles dataset [14]. To that end we selected six configurations (two acoustic models, three musicological models) with the optimal parameters emerging from the experiments on our dataset for a segment and hop size of 11 frames. The performances of these systems are displayed in Table 5 [1]. As one can see, the gain that can be obtained by replacing the uniform musicological model by a trained or theoretical model is again about 2%. Also persisting is the observation that the trained and the theoretical model are very competitive, even though the trained model was trained on another dataset.

S	cosine	Gaussian
7	72.06	74.02
11	72.14	74.10
21	72.26	73.13
31	71.74	71.99

Table 4: AOS in function of segment size S for both acoustic models and the theoretical musicological model with hop size of 1 frame

	cosine	Gaussian
trained	75.36	76.50
theoretical	75.15	76.14
uniform	74.26	74.24

Table 5: AOS for all combinations of acoustic and musicological models run on the Beatles dataset

5. CONCLUSION AND FURTHER WORK

We presented a probabilistic framework for the simultaneous extraction of keys and chords. Its back-

[1] Because we evaluated only at times where one of the basic triads was annotated, these values should not be compared to those from the MIREX contest.

P_s	cosine	Gaussian
0.50	69.90	71.86
0.70	70.89	72.32
0.80	71.49	72.49
0.885	71.72	72.65
0.95	71.39	71.78
0.99	70.94	70.53

(a)

P_d	cosine	Gaussian
0.50	71.35	72.36
0.70	71.53	72.80
0.80	71.72	72.65
0.95	71.67	72.31
0.99	71.67	71.90
1	70.53	71.49

(b)

κ	cosine	Gaussian
0	69.04	68.80
0.25	71.21	71.22
0.4	71.19	71.96
0.5	71.72	72.65
0.6	70.66	72.18
0.75	68.62	70.34

(c)

Table 3: *AOS* in function of P_s (a), P_d (b), κ (c) for both acoustic models and the theoretical musicological model with segment and hop size of 11 frames

end consists of an acoustic model, a duration model and a musicological model. Simultaneously extracting keys and chords allowed us to define a musicological model in terms of chords degrees relative to a key, and as such is only depending on the mode of the key. Through a set of carefully chosen free parameters and multiple options for the acoustic and musicological model, we were able to evaluate the influence of these components independently. Most remarkable was the observation that a uniform musicological model only performs 2% worse than a trained or a theoretically derived model. Also remarkable was that the trained model does not outperform the theoretical model, not even in a test on the dataset it was trained on. The inclusion of a duration model however, leads to a significant gain in *AOS*. The small benefit induced by the musicological model complies with the observation that the relatively simple approach of [3], ignoring all musicological knowledge, was among the best in the MIREX evaluation contest of 2009.

Taking these results in account, our future work will mainly focus on improvements in the front-end and the acoustic model. Another option is to extend the duration model to make it dependent on the chord degree, where a logical musicological assumption would be that the duration of a diatonic chord on the tonic is longer than on other degrees.

6. ACKNOWLEDGEMENTS

This work was conducted in the context of the "Semantic description of musical audio (GOASEMA)" project, which is funded by the "Bijzonder Onderzoeksfonds (BOF)", Ghent University under contract GOA-1250604.

7. REFERENCES

[1] Takuya Fujishima. Realtime chord recognition of musical sound: a system using Common Lisp Music. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 464–467, 1999.

[2] Christopher A. Harte and Mark B. Sandler. Automatic chord identification using a quantised chromagram. In *Proceedings of the 118th Convention of the Audio Engineering Society*, 2005.

[3] Laurent Oudre, Yves Grenier, and Cédric Févotte. Template-based chord recognition: influence of the chord types. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR'09)*, pages 153–158, 2009.

[4] Juan Pablo Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signals. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, pages 304–311, 2005.

[5] Hélène Papadopoulos and Geoffroy Peeters. Large-scale study of chord estimation algorithms based on chroma representation and HMM. In *Proceedings of the 5th International Workshop on Content-Based Multimedia Indexing (CBMI'07)*, pages 53–60, 2007.

[6] Kyogu Lee and Malcolm Slaney. Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio. *IEEE Transactions on Audio,*

Speech and Language Processing, 16(2):291–301, 2008.

[7] Alexander Sheh and Daniel P.W. Ellis. Chord segmentation and recognition using EM-trained hidden Markov models. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR'03)*, pages 183–189, 2003.

[8] Arun Shenoy and Ye Wang. Key, chord, and rhythm tracking of popular music recordings. *Computer Music Journal*, 29(3):75–86, 2005.

[9] Matthias Varewyck, Johan Pauwels, and Jean-Pierre Martens. A novel chroma representation of polyphonic music based on multiple pitch tracking techniques. In *Proceedings of the 16th ACM International Conference on Multimedia (MM'08)*, pages 667–670, 2008.

[10] George Tzanetakis and Perry Cook. MARSYAS: a framework for audio analysis. *Organised sound*, 4(3):169–175, 2000.

[11] Benoit Catteau, Jean-Pierre Martens, and Marc Leman. A probabilistic framework for audio-based tonal key and chord recognition. In *Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V.*, pages 637–644, 2006.

[12] Fred Lerdahl. *Tonal pitch space*. Oxford University Press, New York, 2001.

[13] MIREX audio chord detection task. `http://www.music-ir.org/mirex/2009/index.php/Audio_Chord_Detection_Results`.

[14] Christopher Harte, Mark Sandler, Samer Abdallah, and Emilia Gomez. Symbolic representation of musical chords: a proposed syntax for text annotations. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, 2005.